The Journal of Studies in Language

The Journal of Studies in Language 35(4), 563-582 (2020)

재난재해 소셜 미디어 텍스트 감성분석을 위한 키워드 노이즈 제거 방법론 연구

황창회* · 남지순**

한국외국어대학교

A Methodology of Filtering Irrelevant Keywords for Sentiment Analysis of Disaster-related Social Media Texts

Hwang, Chang-Hoe* and Nam, Jee-Sun**

Hankuk University of Foreign Studies

***First Author / **Corresponding Author**

https://doi.org/10.18627/jslg.35.4.202002.563

pISSN : 1225-4770
eISSN : 2671-6151

Received: January 01, 2020

Revised: January 29, 2020

Accepted: February 11, 2020

ABSTRACT

The Journal of Studies in Language **35.4, 563-582**. This study proposes a methodology for constructing linguistic resources in order to eliminate irrelevant keywords from social media texts related to disasters, such as earthquakes or typhoons. When collecting disaster-related social media texts for sentiment analysis, a large number of noisy keywords metaphorically used, such as 'pupil-earthquake = astonishment,' is observed. In this regard, filtering these linguistic noisy expressions plays a crucial role in performing an accurate text classification or sentiment analysis. In this study, two types of linguistic patterns are examined for filtering noisy expressions in natural & social disaster-related texts, and a bootstrap method based on the DECO Korean electronic dictionary and Local-Grammar Graph(LGG) formalism is suggested. In this way, for six keywords, around 110~ 470 patterns per keyword are described in LGGs. By applying them to a new corpus through the DECO Noise-Filter platform, we obtained about 88.4% f-measure. The methodology suggested in this study may be adopted in filtering other types of noisy expressions, which will improve the reliability of the performance of sentiment analysis of social media texts. **(Hankuk University of Foreign Studies)**

Keywords: Irrelevant keywords, Noise filter, Disaster-related information, Social media text, DECO Electronic Dictionary, Local Grammar Graph

1. 서 론

본 연구는 재난재해와 관련된 사용자 생성문(user-generated sentence)을 수집하는 데에 있어, 키워드의 본래 의미와 무관한 노이즈(noise) 현상을 제거하기 위한

- 본 연구는 2019년도 대한민국 교육부/한국연구재단 지원(NRF-2018S1A5A2A01028628)과 2019년도 한국외국어대학교 교내학술연구지원에 의해 수행되었음. 본 연구의 초기 버전은 인도 IIT-Guwahati 대학과 한국외대 DICORA 연구센터가 공동주최한 IHW-2019 워크샵에서 구두로 발표된 내용에 기초하였다.

전처리(preprocessing) 방법론을 제안하는 것을 목표로 진행되었다. 재난재해 키워드를 통해 SNS (Social Network Service)에서 수집되는 사용자 생성문은 재난재해의 징후 및 피해 상황 예측을 위한 감성분석을 수행하는 데에 있어서 중요한 자원이 된다. 하지만 재난재해 키워드가 SNS 텍스트 상에서 각종 은유 표현(metaphoric expression) 및 신조어(neologism)를 구성하는 요소로써 나타남에 따라 문제가 발생한다. 예문 (1)에서 나타나는 표현들은 이를 잘 보여주는 예시이다.

(1) 재난재해 텍스트를 수집하는 과정에서 나타나는 '노이즈 키워드' 기반 구문의 예
   a. [지진] → 동공 지진이네요 ㅋㅋ 누나도 새해 더 행복하세요
   b. [충돌] → 이번 사건으로 내면에서 엄청 충돌을 겪는 중이예요
   c. [폭풍] → 너무 많은 감정들이 폭풍처럼 할퀴고 지나가
   d. [태풍] → 찻잔 속 태풍이지요

위 예문은 SNS 플랫폼 트위터에서 나타나는 재난재해 단어에 대한 트윗(tweet) 검색결과에서 발췌한 것으로, 예문 (1a)에서 나타나는 "동공 지진"은 '동공이 매우 흔들리며 당황함'을 나타내는 은유적 표현이다. 이는 재난재해 어휘 전·후에 명사가 결합하여 새로운 은유적 표현을 구성하는 신조어 및 합성어(compound) 유형으로, 해당 표현의 경우 재난재해 어휘가 가지고 있는 사건의 양태가 전·후행하는 명사와 연결되어 은유적 의미가 파생되는 것을 알 수 있다. (1b)의 표현 또한 (1a)에서와 유사하게 재난재해 어휘와 명사가 결합하여 실제적 재난과 관련없는 의미를 도출하는 유형이다. 이 경우는 구문 내부에 조사나 부사가 삽입되어 실현되는 것을 볼 수 있다. (1c)와 같이 재난재해 어휘가 부사격 조사 '처럼' 등과 결합하여 '모양이 서로 비슷하거나 같음'을 나타내는 경우, 재난재해 표현의 양상에 따른 비유적 표현을 구성하게 되어 실제 재난재해 관련 텍스트로 분류되기 어렵다. 마지막으로, (1d)에서처럼 어떤 관용구의 일부로써 재난재해 어휘가 포함되어 있는 경우, 실제 재난재해와 동떨어진 표현이 되므로 이 경우는 실제 재해나 재난과 관련 없는 텍스트로 분류되어야 한다.

이와 같이 재난재해 관련 어휘가 실현된 비유적 또는 관용적 표현들이 SNS 상에서 빈번하게 출현하게 되면, 실제 재난재해와 관련된 대상 텍스트를 수집할 때 이러한 '노이즈(noise)' 유형을 필터링(filtering)하는 것이 중요한 작업이 된다. 본 연구에서 2017년 1월 1일부터 12월 31일까지의 트윗 텍스트 중 키워드 "지진"을 포함하고 있는 트윗을 수집한 결과 17,203개가 수집되었는데, 여기서 "동공 지진", "동공 대지진"이 포함된 트윗이 10,148개, 전체의 60%에 달하는 수치를 보였다. 이는 해당 유형의 노이즈들이 제거되지 않은 대용량 텍스트를 바탕으로 재난 현상으로서의 "지진"과 관련된 언어학적 분석 및 텍스트 마이닝(text mining)을 수행하는 경우, 그 신뢰도 및 정확성 측면에서 많은 문제가 발생할 수 있음을 시사한다.

본 연구는 재난재해 코퍼스 수집 시에 나타나는 노이즈를 제거하기 위한 언어지식 기반 패턴화 방법론을 소개하고, 해당 방법론을 통해 구성된 노이즈 제거 자원의 성능 평가를 통해 그 효용성을 검증하였으며, 이를 사용자가 직접 수정 및 적용할 수 있는 DECO Noise Filter 툴(tool)을 제시하여 그 범용성을 확장하고자 한다.

2. 선행 연구

재난재해와 관련된 사용자 생성문 코퍼스의 노이즈 제거와 관련된 연구를 위해, 기존에 연구된 재난재해 관련 코퍼스 기반 연구들을 살펴보면 우선 Matherson(2018)을 들 수 있다. 이 연구에서는 재난재해 상황에 대한 대중의 의사소통을 분석

해내기 위해 뉴질랜드의 Christchurch 지진 이후에 나타난 50만 여개의 트윗에 기반하여 대규모 코퍼스 분석을 수행하였다. 이를 토대로 재난재해 이후 트위터와 같은 제도적 규범성이 낮은 플랫폼에서 나타나는 담화의 양상을 "공공 규범", "트위터의 행위 유도성", "재난", "지역의 정치 문화" 4가지로 분류하는 연구를 수행하였다. 하지만, 트윗 수집과 관련하여 일부 색인어 및 해시태그를 제공했을 뿐 수집 과정에서 발생하는 노이즈를 어떻게 처리하였는지에 대해서는 언급된 바가 없다.

재난재해와 관련된 또다른 연구로는 동일본 대지진 시에 업로드 된 트위터 사용자 생성문을 바탕으로 대중의 사회 환경적 불안도를 측정하기 위한 방법론을 제시하였던 Baek et al.(2013)의 연구가 있다. 해당 연구는 수집된 트윗 데이터를 주석 코퍼스로 변환하여, 재난과 관련된 토픽 및 공기관계를 추출하는 것을 통해 사회적 불안도 감지에 대한 언어적 데이터를 분류해내는 연구를 수행하였다. 하지만 마찬가지로, 데이터 습득 및 구성을 포함하여 해당 데이터에서 대상 데이터만을 추출하기 위한 노이즈의 제거 과정은 언급되어 있지 않다. 제시된 연구들 이외에도 재난재해와 관련된 국내의 SNS 코퍼스 연구들은 노이즈 제거와 관련된 체계적인 방법을 제시하고 있는 경우를 찾아보기 힘들다.

본고에서는 한국어에서 나타나는 재난재해 어휘를 선별해내기 위해 신자행(2016), 박태연 외(2017)에서 제시된 국가재난정보관리시스템(National Disaster Management System: NDMS) 리스트의 일부를 참고하였다. 해당 리스트는 각 기관별로 보유하고 있는 재난에 관련된 정보를 통합함에 따라 재난 발생 시 피해를 최소화하는 것을 목적으로 제작된 시스템 내에 포함되어 있는 것으로, 재난의 양상을 자연재해 및 사회재난의 대분류로 분류하고 있어 본 연구를 진행하기 위한 코퍼스 데이터의 대상 어휘를 선정하는 데에 기준점으로써 활용하였다.

3. 재난재해 코퍼스에서의 노이즈 추출 방법론

3.1 데이터 수집

본 연구에서는 SNS 데이터 수집을 위해 자연재해와 사회재난에 직접적으로 관련되는 키워드를 각 3개씩 선정하였다. 선정된 어휘들이 나타나는 사용자 생성문을 수집하기 위해서 한국외국어대학교 DICORA 연구센터에서 제공하는 Deco T-Crawler(황창회 · 남지순, 2018)를 활용하였다.

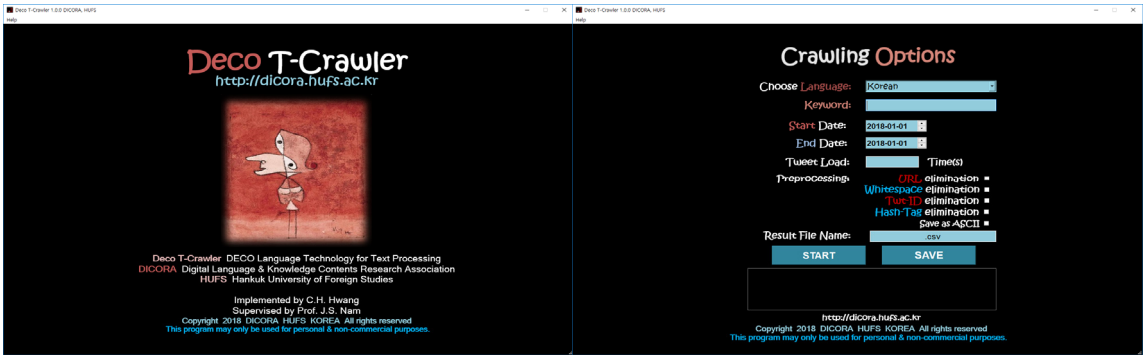

그림 1. 트위터 크롤링 프로그램 Deco T-Crawler

Deco T-Crawler는 색인어에 기반하여 해당 기간에 작성되었던 트윗의 목록을 자동으로 수집한 후, 기본적인 전처리(preprocess) 작업을 사용자가 설정한 옵션에 따라 적용하여 출력해주는 트위터 크롤링 프로그램이다. 본고에서는 해당

프로그램을 활용하여 2017년 한 해 동안 대상 재난재해 어휘가 출현한 트윗을 트위터 웹 페이지에서 수집하였다. 이에 대한 구체적인 트윗 텍스트의 개수 및 그 어절 정보는 다음 표와 같다.

표 1. 수집된 자연재해/사회재난 코퍼스의 트윗 및 어절 개수

자연 재해 분야			사회 재난 분야		
재난 키워드	수집된 트윗 개수	어절 개수	재난 키워드	수집된 트윗 개수	어절 개수
가뭄	17,884	278,048	붕괴	20,291	346,282
지진	21,183	317,313	충돌	19,583	356,914
홍수	26,491	443,441	폭발	24,631	406,416

3.2 노이즈 추출 관련 도구 및 연구 방법론

재난재해 코퍼스에서 나타나는 노이즈 표현들을 추출해내기 위해 본 연구에서는 DECO 한국어 기계가독형 전자사전(남지순, 2018)을 토대로 모든 대상 코퍼스의 형태소분석을 수행하였다. DECO 사전은 한국어의 체언 및 용언에 대한 표제어와 그 활용형에 대한 어휘·문법 정보를 제공하는 대용량 전자사전이다. 해당 사전에 수록되어 있는 정보들은 유한상태 트랜스듀서(Finite State Transducer: FST)로 구조화되어 UNITEX 플랫폼(Paumier, 2003)에서 코퍼스에 적용된다.

사전이 적용된 코퍼스에 기반하여 재난재해 어휘가 포함되어 있는 노이즈 용례들을 일괄적으로 확보하기 위해, 본고에서는 부분문법 그래프(Local Grammar Graph: LGG) 모델을 활용하였다. LGG는 프랑스의 전산언어학자 모리스 그로스(Maurice Gross)에 의해 제안된 언어 처리 모델로(Gross, 1997; 1999), 텍스트에서 나타나는 국지적인 문법 현상들을 기술하고 처리하는 데에 최적화되어 있는 언어 기술 방법론이다. 본 연구에서는 DECO 사전의 품사/활용형 태그를 활용하여 노이즈 표현의 형식에 대한 구체적인 패턴을 LGG 형식으로 구축하였으며, UNITEX 플랫폼을 통해 이를 코퍼스에 적용함으로써 해당 패턴을 코퍼스 전체에 걸쳐 검색하고, 그 노이즈 유형을 일괄 도출하였다.

Unitex 플랫폼에서 LGG를 적용하여 추출된 어휘 패턴의 기술 통계적 정보를 바탕으로 재해재난 어휘와 관련된 노이즈 현상을 효율적으로 살펴보기 위해 대상 어휘와 인접 어휘 간의 거리를 벡터공간(vector space) 내에 유사도에 따라 배치시켜 클러스터링(clustering)하는 Word2Vec 모듈을 사용하였다. 이를 위해 DecoTex(유광훈 · 남지순, 2017) 프로그램의 {Expanding Lexica via Word2Vec} 모듈을 사용하였으며, 해당 모듈을 기반으로 재해재난 어휘와 유사성을 가진 어휘들을 가까운 순서대로 추출하였다. 이에 따라 도출되는 빈도 통계정보 및 Word2Vec 도출 결과를 교차 비교함에 따라 노이즈 표현을 분류하였다.

이러한 과정을 거쳐 실제적으로 재난재해 코퍼스에서 나타나는 노이즈 구문을 추출하기 위해, 대상 코퍼스에서 관찰되는 노이즈 표현들을 몇 가지 구조적 특징을 기반으로 분류하였다. 그 중 첫 번째 노이즈 유형은 재난재해 어휘와 일반 명사의 결합으로 구성되는 표현으로, 예를 들어 '동공 지진', '폭풍(으로) 흡입'과 같은 유형이 여기 해당한다. 이들은 코퍼스 내에서 다음과 같은 구조로 나타난다.

☞ **재난재해 어휘 + (조사) + (부사) + 일반 명사**
☞ **일반 명사 + (조사) + (부사) + 재난재해 어휘**

해당 표현에 실현되는 '명사' 부류를 추출해내기 위해, 본 연구에서는 위와 같은 명사 연쇄를 인식할 수 있도록 LGG를 구성하여 이를 코퍼스에 적용하였다. 이를 통해 코퍼스에서 관련 명사 연쇄({NN}) 유형을 포착하고, 재난재해 어휘 좌우로 결합할 수 있는 명사의 리스트를 빈도순으로 확인하였다. 이러한 과정을 통해 수집된 명사들은 이후 해당 유형의 재난재해 관련 노이즈 표현을 일괄 추출하기 위한 패턴 문법 기술에 활용되었다.

두 번째는 재난재해 어휘에'-처럼/같이'과 같이 어떤 모양이나 특징을 표현하는 부사격 조사가 동반되는 비유 표현의 유형이다. 예를 들어 '찻잔 속의 태풍처럼'이나 '폭풍처럼 지나간 학창시절'과 같은 표현이 여기 해당된다. 이 경우 재난재해 어휘를 비유적으로 환원하는 일련의 비교격 조사 유형의 출현과 또한 좌우에 실현되는 용언 부류와의 공기 관계({NV}) 현상을 파악하는 것이 필요하다. 이러한 유형을 추출해내기 위해 아래와 같은 결합 관계 패턴을 기술하여 이를 코퍼스 내에서 추출해 내었다.

☞ **재난재해 어휘 + 비교격 조사 + 용언 활용형**
☞ **용언 활용형 + 재난재해 어휘 + 비교격 조사**

앞서와 마찬가지로, 해당 유형이 코퍼스 내에서 어떻게 실현되는지 살펴보기 위해 DECO 사전이 적용된 코퍼스에 LGG로 구성된 노이즈 유형 패턴을 적용하였으며, 해당 과정에서 도출된 빈도 정보를 바탕으로 노이즈 표현에 해당하는 체언 및 용언 리스트를 구축하여 노이즈 패턴을 구성하였다.

이에 부가적으로, 재난재해 어휘가 포함되어 있는 속담 및 관용구를 노이즈 리스트에 추가하기 위해 본 연구에서는 네이버 관용구 사전[1]에 등재되어 있는 목록을 활용하였다. 해당 목록은 객체 지향 프로그래밍 언어인 Python의 라이브러리인 BeautifulSoup를 통해 자동으로 추출된 것으로, 네이버 관용구 사전 내 840여개의 관용구 목록 중 재난재해 대상어휘가 관용구 구성상에 직접 나타나는 표현만을 추출하여 이를 토대로 노이즈 리스트를 확장하였다.

4. 노이즈 처리를 위한 리스트 구축

본 연구에서 사용하는 '자연재해'와 '사회재난'은 다음과 같은 기준에 의해 분류된다. 실제로 재해와 재난에 대한 분류는 여러 가지 현상적인 측면에서 논의되어 이론에 따라 많은 분류체계가 제시되어 있지만, 여기서는 이에 대한 정의를 현행 재난안전법 상에서 제안하는 자연재해 및 사회재난의 분류체계를 따르기로 한다.

(2) 재난안전법 상에서 나타나는 자연재해 사회재난의 분류
   a. 자연재해: 가뭄, 강풍, 낙뢰, 대설, 자연우주물체의 추락·충돌, 조류대 발생, 조수, 지진, 태풍, 풍랑, 해일, 호우, 홍수, 화산활동, 황사 등
   b. 사회재난: 가축전염병, 감염병, 교통사고, 항공사고, 해상사고, 국가기반체계의 마비, 붕괴, 폭발, 화생방사고, 화재, 환경오염사고, 해외재난 등

1) https://ko.dict.naver.com/#/topic/search?category1=idiom

위의 기준은 앞서 2장에서 언급한 NDMS 시스템 상에 세분화되어 적용된 바 있다. 본 연구에서는 이러한 자연재해/사회재난 유형에서 {가뭄/태풍/홍수}와 {붕괴/충돌/폭발} 키워드를 추출하여 코퍼스 수집을 진행하였다. 특히 키워드 {충돌}은 각종 교통, 항공, 해상에서 발생할 수 있는 세부적인 재난을 상정한 것으로, 사회재난 도메인을 대표적으로 표현하는 중요한 키워드의 하나로 판단된다.

본 연구에서는 100만 어절 규모의 자연재해 {가뭄/지진/홍수}와 사회재난 {붕괴/충돌/폭발} 관련 트위터 텍스트를 추출하여 노이즈 유형을 분석하였다. 이 장에서는 각 재난재해 유형에 따른 노이즈 처리 리스트의 추출 및 자원 구축 과정을 자세히 살펴보기로 한다.

4.1 자연재해 {가뭄/지진/홍수}에 대한 노이즈 처리 리스트 구축

4.1.1 {NN} 유형의 노이즈 패턴

4.1.1.1 자연재해 어휘와 공기하는 명사 유형 추출을 위한 LGG 구축

자연재해 어휘와 명사의 결합 유형을 추출하여 노이즈 리스트로 구축하기 위해, 본고에서는 사전이 적용된 자연재해 코퍼스에 "재난재해 어휘 + (조사) + (부사) + 일반 명사" 및 "일반 명사 +(조사) + (부사) +재난재해 어휘" 유형을 인식하기 위한 LGG를 적용하였다. 해당 LGG의 구성은 다음과 같다.

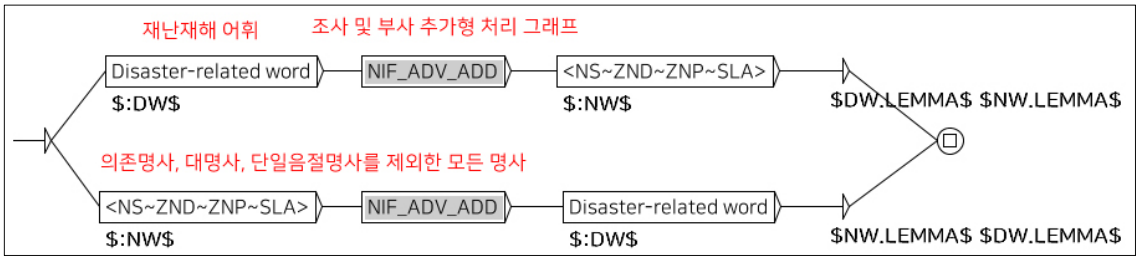

그림 2. 자연재해 어휘와 결합하는 명사 유형을 추출하기 위한 LGG

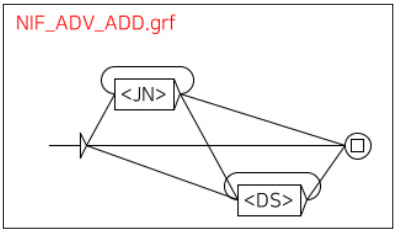

그림 3. 명사 곡용 및 부사 삽입을 처리하기 위해 그림 2에 삽입된 서브그래프 {NIF_ADV_ADD.grf}

<그림 2>의 상단 경로는 자연재해 어휘가 선행한 후 명사가 오는 패턴을 인식하며, 하단 경로는 일반 명사 어휘가 선행한 후 자연재해 어휘가 오는 경우를 나타낸다. 해당 그래프에서 어휘 및 품사 부류를 표상하는 각각의 박스 중 {Disaster-related word}의 자리에는 각 코퍼스 도메인에 따른 자연재난 어휘들이 위치하게 되며, <NS~ZND~ZNP~SLA>의 경우 DECO 사전의 태그 목록(tag-set)에 따라 명사부류(NS) 중 의존명사(ZND) 및 대명사(ZNP)와 1음절 명사(SLA)를 '제외한('~'표시

함수의 의미)' 명사 부류를 나타낸다. 이때, 두 명사의 결합 관계 사이에서 나타날 수 있는 조사 및 부사는 <그림 3>과 같은 서브그래프 {NIF_ADV_ADD.grf}를 통해 처리되는데, 해당 그래프에서의 <JN>과 <DS> 태그는 각각 DECO 사전에서의 조사 및 부사를 의미한다.

본 과정에서는 재난재해 어휘와 결합하는 명사의 유형을 인식하기 위해 재난재해 어휘 및 이와 결합하는 명사를 각각 $:DW$, $:NW$의 변수로 함수화한 후, 각각에 $DW.LEMMA$와 $NW.LEMMA$의 연산자를 사용하여 레마(lemma) 형태만을 출력해주는 방식으로 구현하였다. 위 과정을 통해 추출된 {가뭄/지진/홍수} 코퍼스에서 나타나는 자연재해 어휘와 명사의 결합 유형의 일부를 보이면 <표 2>와 같다.

표 2. {가뭄/홍수/지진} 코퍼스의 자연재해 어휘와 명사 결합 유형에 대한 상위 빈도 리스트

가뭄		홍수		지진	
일치 유형	빈도	일치 유형	빈도	일치 유형	빈도
떡밥 가뭄	250	떡밥 홍수	1160	동공 지진	13040
가뭄 극복	159	노아 홍수	987	포항 지진	462
가뭄 대책	143	홍수 미스터리	722	지진 발생	372
가뭄 심각	95	눈물 홍수	579	경주 지진	177
가뭄 고생	90	홍수 피해	494	지진 피해	174
가뭄 피해	87	홍수 예방	338	지진 관련	93
가뭄 이후	66	가뭄 홍수	300	일본 지진	73
연성 가뭄	57	정보 홍수	242	기준 지진	64
가뭄 걱정	55	홍수 영업질	235	지진 동공	61
가뭄 대비	50	홍수 이후	224	멕시코 지진	57
최악 가뭄	46	홍수 방지	211	오늘 지진	52
가뭄 해소	40	내야 홍수	211	지진 주의	44
가뭄 농작물	39	사람 홍수	209	카메라 지진	42
요즘 가뭄	33	홍수 위험	203	지진 해일	40
아내 가뭄	32	홍수 문제	180	이번 지진	40
가뭄 가운데	31	정비사업 홍수	162	이상 지진	36
가뭄 현장	31	홍수 순간	123	규모 지진	30
가뭄 단비	30	두려움 홍수	101	지진 오늘	29
가뭄 홍수	29	홍수 버티기	96	지진 원전	24
가뭄 해결	28	홍수 인해	94	북한 지진	24

이를 자세히 살펴보면, 세 가지 자연재해 어휘에 대한 코퍼스에서 자연재해 어휘와 명사의 결합으로 이루어지는 노이즈 유형에는 비유적 표현 및 신조어를 비롯하여, 트윗 작성 중 발생하는 오타나 동형이의어(homograph) 형태 등이 포함된다. 이들 중 가장 상위 빈도를 차지하는 표현들은 모두 노이즈 표현에 해당하는 "떡밥 가뭄 / 노아 홍수 / 동공 지진" 이었으며, {가뭄} 도메인에서 노이즈가 상대적으로 적게 출현한 것을 확인할 수 있다. {홍수} 도메인과 같은 경우 노이즈의 양상이 {떡밥 홍수 / 홍수 미스터리 / 쇼핑몰 홍수} 등으로 다양하게 나타나는 반면, {지진} 도메인과 같은 경우 최상위 빈도에 해당하는 "동공 지진"이 차상위에 위치해있는 "포항 지진"의 빈도에 약 28배에 해당하고 있음을 확인할 수 있다.

이와 같은 빈도 탐색 결과는 자연재해 어휘와 공기하는 명사들을 일괄 추출한 것으로, 본 연구에서 구축하고자 하는 노이즈 명사 목록의 기초 데이터의 역할을 한다. 다만 이와 같은 단순 공기 빈도에 기반하는 통계 결과의 한계를 보완

하기 위해 본 연구에서는 DecoTex(유광훈· 남지순, 2017) 프로그램의 {Expanding Lexica via Word2Vec} 모듈을 활용하여 통계적으로 획득된 유사어휘 후보군을 2차적으로 활용하는 방식을 채택하였다. 다음은 DecoTex 프로그램의 해당 모듈을 보인다.

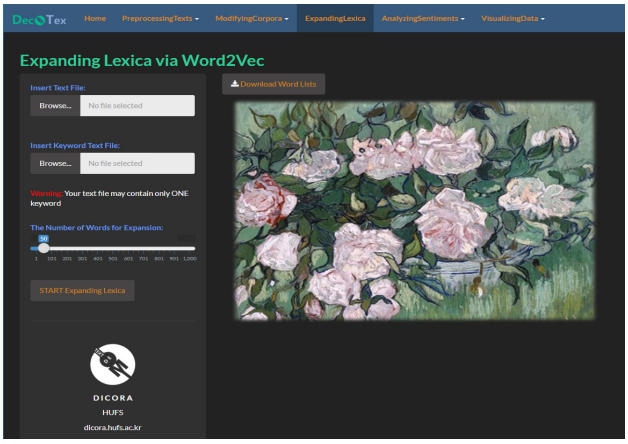

그림 4. DecoTex 프로그램의 {Expanding Lexica} 모듈

이 모듈에서는 사용자가 입력한 특정 키워드에 대하여 어휘 유사도를 측정하여 주변 어휘들을 클러스터링 및 그 결과값을 산출해준다. 대상 어휘가 코퍼스 상에서 어떠한 어휘와 공기(co-occur)하는지를 계산하여 이를 토대로 궁극적으로는 대상 어휘와 관련된 어휘 사전의 확장(Lexicon expansion)을 수행하기 위한 데이터를 제공한다.

본 연구에서 {Expanding Lexical}을 통해 산출한 {가뭄/홍수/지진} 코퍼스 내 자연재해 어휘의 Word2Vec 결과 일부를 보이면 다음과 같다. 기준 어휘의 벡터 값에 기준하여 0과 가까운 어휘일수록 대상 어휘와 가까운 어휘로 볼 수 있다.

(3) 자연재해 도메인 어휘에 대한 Word2Vec 결과-1: 상대적으로 연관성을 보이는 어휘 부류

a. {가뭄}과 연관성이 있는 것으로 제시된 어휘류: 극복(0.48), 대응(0.51), 발표(0.53), 예·경보(0.53), 대책(0.53) 등

b. {홍수}와 연관성이 있는 것으로 제시된 어휘류: 개선(0.55), 경보(0.56), 생태환경(0.58), 지난번(0.58) 등

c. {지진}과 연관성이 있는 것으로 제시된 어휘류: 현장(0.47), 답변(0.48), 주말(0.49), 이유(0.5) 등

(3)의 목록을 자세히 살펴보면, 먼저 {가뭄}과 같은 경우 자연현상으로써의 {가뭄}을 극복 및 타개하기 위한 "극복", "대응"과 같은 명사 어절들이 주를 이루었고, "발표, 예·경보"와 같이 가뭄 정보의 전파와 관련되어 사용될 수 있는 어절들이 비교적 높은 연관성을 가진 것으로 나타나는 것을 볼 수 있다. {홍수} 도메인의 경우도 마찬가지로 "개선", "경보"와 같이 재난 상황에 대한 방안 및 전달과 관련된 어절이 높은 유사도로 나타났으며, {지진} 도메인의 경우 재해와 관련된 "현장" 또는 정부 방침과 관련되어 있는 "답변" 등의 명사들이 높은 벡터값으로 나타나고 있었으며, 해당 재해가 발생된 시간 및 날짜에 대한 표현들이 고유사도(highly approximated) 표현으로 나타나는 것을 확인할 수 있다.

하지만 이러한 통계적 추정 결과는 사용된 확률 알고리즘뿐 아니라 계산을 위한 데이터의 질과 양에 큰 영향을 받기 때문에 전적으로 신뢰하기 어렵다. 가령 아래 (4)의 목록에서와 같은 어절들 또한 높은 유사도를 가진 것으로 나타나는데, 이를 통해 개별 도메인의 Word2Vec 결과에 대한 내용이 개별적으로 검토되어야 할 필요성이 확인된다.

(4) 자연재해 도메인 어휘에 대한 Word2Vec 결과-2: 실제 어휘적 연관성이 적은 부류

a. {가뭄}과 연관성이 있는 것으로 제시된 어휘류: 끝에(0.52), 포함(0.54), 좋아해(0.54), 운영(0.55) 등

b. {홍수}와 연관성이 있는 것으로 제시된 어휘류: 망라한(0.55), 멀티사업입니다(0.56), 속에서(0.58), 덕에(0.59) 등

c. {지진}과 연관성이 있는 것으로 제시된 어휘류: 찾아(0.44), 지진짤(0.492), 잡은(0.5), 학부모(0.51) 등

위에 제시된 어절들은 앞선 목록들과 달리 재해 도메인 코퍼스 내에서 키워드들과 고유사도를 보이는 어절들로 판단하기 어렵다. 이러한 관찰을 통해 이러한 자료들을 그 자체로 노이즈 제거를 위한 패턴으로 간주하는 것은 적절치 않음을 확인할 수 있다. 그러나 실제 개별적으로 구축되어야 하는 패턴 문법의 중요한 토대가 되는 어휘 정보를 효과적으로 제공한다는 장점이 있다.

4.1.1.2 공기정보 및 Word2Vec을 통해 추출된 명사에 기반한 {NN} 유형의 노이즈 표상 LGG

이상과 같은 과정을 통해 자연재해 어휘 {가뭄/홍수/지진}의 좌우에 공기하는 명사 리스트를 수집하였으며, 이를 기반으로 각각의 코퍼스에서 나타나는 {NN} 유형의 노이즈 리스트를 빈도순으로 추출하였다. 또한 Word2Vec에서 추출된 유사도 높은 명사들의 리스트를 점검하여 이를 기반으로 현재 {NN} 유형의 노이즈 리스트에 포함되어야 하는 명사 부류를 확장하였다. 이렇게 획득된 노이즈 명사의 리스트 일부를 제시하면 다음과 같다.

(5) {가뭄/홍수/지진} 코퍼스의 자연재해 어휘와 명사 결합 유형 노이즈 리스트 일부

a. {가뭄}을 포함하는 {NN}유형의 노이즈 예: 떡밥-가뭄, 사람-가뭄 등

b. {홍수}를 포함하는 {NN}유형의 노이즈 예: 떡밥-홍수, 눈물-홍수, 정보-홍수, 홍수-시대 등

c. {지진}을 포함하는 {NN}유형의 노이즈 예: 동공-지진, 화면-지진 등

분류된 노이즈 명사들 살펴보면, 먼저 자연재해 도메인 {가뭄}과 {홍수}에서는 "흥미를 끌만한 사건 등이 적다/많다"의 의미로 사용되는 "떡밥-가뭄, 떡밥-홍수"와 같은 표현들이 매우 높은 빈도로 나타나는 노이즈 표현임을 볼 수 있다. {지진}코퍼스에서는 "어떤 상황 및 사건에 의해 몹시 당황하다"의 뜻을 가지고 있는 "동공-지진"이라는 표현이 매우 높은 빈도로 나타나는 것을 확인할 수 있다.

노이즈 명사가 가지는 언어적 특징은 비교적 명백한데, 이는 해당 재해 어휘가 가지고 있는 어떠한 서술적 양상, 즉 "가뭄"과 같은 어휘에서 "마르다/건조하다"와 같은 어휘적 특성을 추출하여 전·후치되는 명사에 부여함에 따라 은유성 표현들을 만들어내는 것이다. 따라서 "분량 가뭄"과 같은 명사구는 "분량이 가뭄과 같이 메말랐다", 즉 "분량이 부족하다"라는 뜻으로 사용되며, 신체 관련 명사들이 부착되는 경우 "해당 신체부위가 매우 건조함"을 의미하게 되고, 추상 명사와 결합하는 "마음 가뭄"과 같은 때는 "마음이 풍족하지 못함"을 의미하는 은유적 표현이 구성된다. 또한 재해재난 어휘들이 어떠한 동적 양상을 나타내는 용도로 활용됨에 따라 노이즈 명사구 구성에서 뒤에 위치하는 경향을 보이는데, 때문에 노이즈를 구성하는 재해 어휘 전치 명사의 비율이 후치 명사에 비해 높게 나타나는 특성을 가지고 있으며, 이러한 특징은 모든 재난재해 도메인에 동일하게 나타나고 있다.

이상의 관찰결과를 토대로 구축된 리스트를 이용하여 실제 코퍼스에서 나타나는 이러한 노이즈 유형을 인식하고 추출하기 위해 다음과 같은 LGG를 구축하였다.

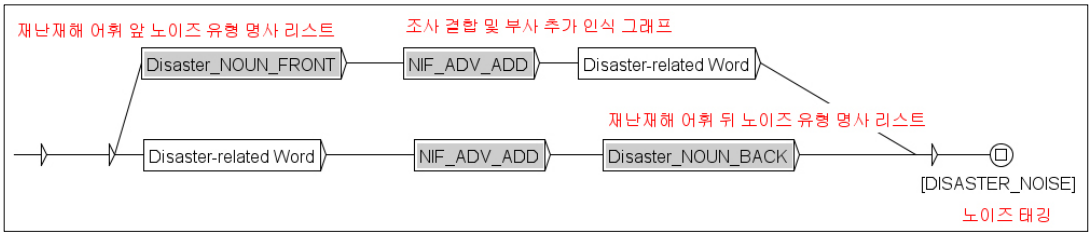

그림 5. 재난재해 어휘와 명사가 결합한 노이즈 패턴을 표상하는 LGG

재난재해 텍스트에서 나타나는 노이즈 표현들을 표상하는 그래프는 총 세가지 하위 그래프의 구성으로 이루어진다. 여기에 삽입된 {재난재해_NOUN_FRONT.grf}와 {재난재해_NOUN_BACK.grf} 서브 그래프는 각각 앞서 재난재해 대상 어휘와 공기하는 명사 리스트를 통해 수집된 전·후치 노이즈 명사가 수록되어 있는 리스트로, 차후 유지 보수에 용이하도록 한글 내림차순으로 구성되었다. 해당 그래프는 앞서 구성한 노이즈 명사 리스트에 더불어 <그림 3>에서 소개한 {NIF_ADV_ADD} 서브 그래프를 통해 단순 명사의 결합형을 포함하여 그 사이에 실현될 수 있는 조사 및 부사의 결합 유형을 일괄적으로 처리해준다. 처리의 마지막 단계에서 추출된 노이즈 유형들에 "[DISASTER_NOISE]" 형태의 노이즈 태그를 부착하도록 구성하여 이후 5장에서 소개될 재난재해 필터(filter) 프로그램을 통해 해당 태그가 부착된 문장들을 일괄 삭제할 수 있도록 구성하였다. 이 그래프를 UNITEX 플랫폼을 통해 코퍼스에 적용하면 <그림 6>과 같이 주석 코퍼스(annotated corpus)를 결과로 획득할 수 있다.

그림 6. {Disaster_Process_지진.grf} 그래프의 {NN} 경로를 통해 추출된 지진 도메인의 노이즈 텍스트

위의 결과에서 보듯이 지진 도메인의 코퍼스에서 {NN} 노이즈 유형을 태깅하는 LGG는 "눈/눈빛/눈동자 지진" 등으로 구성된 노이즈 유형들을 잘 인식하고 있다. 단순 명사 연쇄 구성만을 포착하는 것이 아니라, "눈동자는 그야말로 지진"과 같이 노이즈 명사구 구성에서 나타날 수 있는 조사의 곡용 및 부사 삽입 형태들을 처리하고 있기 때문에 그 활용도가 더욱 높다고 할 수 있다. 지진 도메인에서 인식된 이러한 노이즈 유형은 재해 관련 텍스트를 분석할 때 위에서처럼 제거 대상으로 분류되어야 한다.

4.1.2 {NV} 유형의 노이즈 패턴

4.1.2.1 '비교격 조사를 동반한 자연재해 어휘'와 공기하는 용언 추출을 위한 LGG 구축

자연재해 어휘가 비교격 조사와 결합하는 경우, 일련의 용언과 공기할 때 그 용언의 의미 자질을 강조하기 위해 '지진처럼' 또는 '폭풍처럼' 같은 비유적 부사어처럼 사용되는 경우들이 여기 해당한다. 우선 이러한 자연재해 어휘와 공기하는 용언들을 추출하기 위해서 앞서 {NN} 패턴의 경우와 마찬가지로 다음과 같은 LGG를 구성하였다.

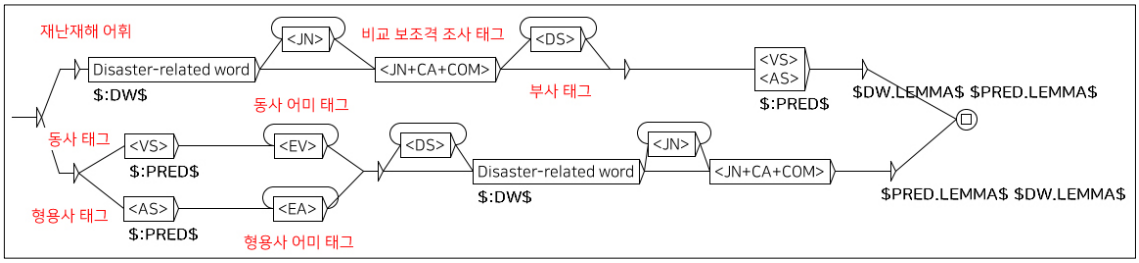

그림 7. 자연재해 코퍼스의 {재난재해 어휘+비교격 조사}와 결합하는 용언 포착을 위한 LGG

앞서와 같이 해당 LGG의 {Disaster_related word} 박스 내에는 자연재해 코퍼스의 대상 어휘인 {가뭄/홍수/지진}이 도메인에 따라 삽입되며, 이와 결합하는 보조격 조사를 인식하기 위해 DECO 사전의 태그셋(DECO-Tagset)에서 제공하는 보조·비교격 조사 태그인 <JN+CA+COM>을 활용하였다. 실제로 이렇게 기술된 구문이 일괄적으로 모두 노이즈 유형에 해당하지 않으므로 재난재해 명사구 좌우에 출현할 수 있는 용언 부류들을 추출하여 개별 검증을 거쳐야 할 필요가 제기되기 때문이다. 이를 위해 DECO사전의 동사 태그 <VS>와 형용사 태그 <AS>를 활용하였다.

또한 그래프의 하단 경로에서는 용언 부류가 대상 어휘에 전치되는 경우를 추출하기 위해 그 위치에 나타나는 용언 활용형들을 고려하였다. 동사와 형용사의 어미 활용형인 <EV>, <EA> 태그가 복합적으로 실현되는 현상을 인식하기 위해 루프를 통해 그 경로를 기술하였다. 또한 자연재해 명사구와 용언부의 결합 사이에서 나타날 수 있는 부사 표현들을 고려하기 위해 <DS> 태그가 사용되었다. 해당 그래프에도 마찬가지로, 어휘 결합의 탐색을 조금 더 용이하게 수행하기 위해 포착된 표현을 기본형(lemma)의 결합쌍으로 변환하는 과정을 거쳤다. 앞서 4.1.1장에서 언급된 것과 같이 재난재해 어휘와 이에 결합하는 용언 부류를 $:DW/PRED$ 연산자를 이용하여 함수화하였고, 이를 각각 $DW.LEMMA$, $PRED.LEMMA$로 출력하는 방식으로 이루어졌다. 이러한 구성의 그래프를 각 자연재해 도메인 코퍼스에 적용해 추출해낸 LGG 패턴의 적용 결과는 다음과 같다.

표 3. {가뭄/홍수/지진} 코퍼스의 자연재해 어휘+보조격 조사와 공기하는 용언의 상위 빈도 리스트

Index	{가뭄} corpus	{홍수} corpus	{지진} corpus
1	가뭄 메마르다	홍수 사랑하다	지진 해이다
2	가뭄 도움	홍수 밀려오다	난폭하다 지진
3	속다 가뭄	홍수 쏟아지다	지진 흔들리다
4	가뭄 말라붙다	홍수 불리다	지진 나다
5	안되다 가뭄	홍수 흐르다	발생하다 지진
6	가뭄 갈라지다	홍수 터지다	시작하다 지진
7	가뭄 식어버리다	홍수 나다	크다 지진
8	가뭄 갈라지다	홍수 넘치다	싶다 지진
9	모르다 가뭄	홍수 범람하다	지진 떨어지다
10	가뭄 이야기하다	홍수 흘리다	무섭다 지진
11	가뭄 무미건조하다	홍수 오다	지진 무섭다
12	불타다 가뭄	홍수 밀려오다	지진 상관없다
13	가뭄 마르다	홍수 타다	지진 비슷하다
14	가뭄 갈라지다	무섭다 홍수	절다 지진
15	가뭄 크다	홍수 발매되다	지진 다가오다
16	가뭄 몰려오다	홍수 튀어버리다	움직이다 지진
17	가뭄 갈라지다	홍수 쌓이다	지진 살아남다
18	메마르다 가뭄	홍수 떠오르다	돌리다 지진
19	가뭄 말라가다	홍수 뜨다	지진 피하다
20	쏟아지다 가뭄	홍수 정리되다	지진 무너지다

해당 용례 추출 결과를 자세히 살펴보면, 자연재해 대상 어휘에 비교격 조사가 결합하는 경우 선·후행하는 체언 및 용언 부류들 또한 해당 은유적 표현을 구성하는 일부로 등장하는 것을 알 수 있다. 예를 들어 대상어휘 "가뭄+{비교급조사}"과 관련된 표현의 경우 "메마르다/식다/말라붙다/갈라지다" 등의 용언이 공기하여 은유적 표현을 구성하므로 이들을 노이즈 표현으로 처리해야할 필요성이 나타난다. 자연재해 어휘 "홍수"에서 나타나는 패턴 또한 이와 유사한 형태로, "밀려오다/쏟아지다/흐르다"등의 용언들이 "홍수+{비교급조사}"와 공기하는 경우 개인의 감정과 관련된 것들을 홍수의 속성에 빗대어 표현하는 은유의 일부임을 볼 수 있으며, "지진+{비교급조사}"과 관련해서는 "흔들리다/난폭하다/무섭다" 등의 용언들이 비유적 용도로 실현되는 것을 관찰할 수 있다.

4.1.2.2 공기정보를 통해 추출된 용언에 기반한 {NV} 유형의 노이즈 표상 LGG

이러한 과정을 통해 본 연구에서는 비교격 조사를 동반한 자연재해 명사구와 용언이 결합한 {NV} 유형의 노이즈 리스트를 다음과 같은 LGG 유형으로 기술하였다.

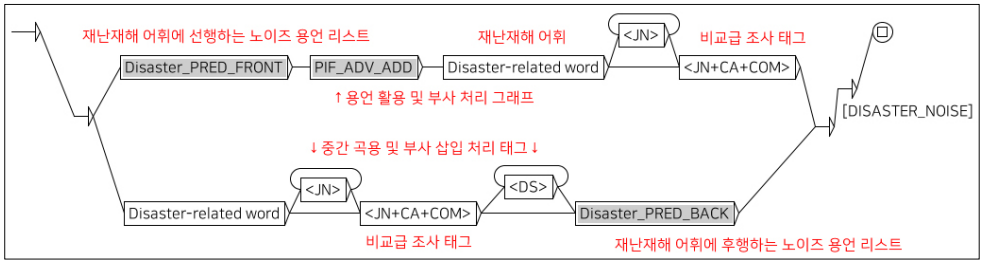

그림 8. 자연재해 코퍼스의 "재난재해 어휘 + 비교격 조사" 결합 유형 포착을 위한 LGG

그래프 경로상의 {Disaster_PRED_FRONT/BACK.grf} 서브 그래프는 비교격 조사와 결합하는 자연재해 명사구에 전·후치되어 나타나는 노이즈 용언들의 리스트로써, 그 내부는 동사 및 형용사가 내림차순 형태로 구성되어 있다. 특히 상단 경로에서 나타나는 재해재난 명사구에 선행되는 용언부의 포착을 위해서는 어미 결합 및 부사 삽입에 대한 특징이 반드시 고려되어야 한다. 위 그래프는 {PIF_ADV_ADD.grf} 서브 그래프를 포함하고 있는데, 여기에는 동사와 형용사에 대한 활용형을 처리하기 위한 '동사 활용어미(<EV>)' 태그와 '형용사 활용어미(<EA>)' 태그 및 부사 처리를 위한 <DS> 태그가 다음 <그림 9>와 같은 방식으로 기술되어 있다.

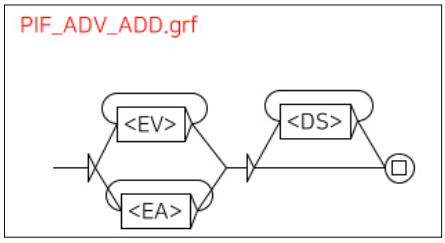

그림 9. 용언의 활용 및 부사 삽입을 처리하기 위한 {PIF_ADV_ADD.grf}

<그림 9>의 그래프에는 재난재해 어휘와 결합하는 비교급 조사들을 나타내는 <JN+CA+COM> 태그와 명사의 조사 결합형을 고려하기 위한 <JN> 태그가 사용되었으며, 부사적 요소 삽입을 처리하기 위한 <DS> 태그가 기술되어 광범위한 영역에서 노이즈 표현들을 처리할 수 있게 하였다. 그래프의 마지막 부분에서는 자연재해 코퍼스에서 포착된 용언 결합형 노이즈 표현들에 "[DISASTER_NOISE]" 태그를 부착할 수 있게 하여, 뒤에서 언급될 프로그램에 적용되어 노이즈를 일괄 처리하는 것을 가능하게 한다.

이러한 과정을 통해 추출된 용언들을 바탕으로 분류한 자연재해 코퍼스의 {NV} 유형 노이즈의 용언부류는 다음에서 보이는 바와 같다.

(6) {가뭄/홍수/지진} 코퍼스의 '자연재해 어휘+비교격 조사'와 용언 결합 노이즈 리스트 일부
   a. {가뭄}을 포함하는 {NV}유형의 노이즈 예: 가뭄-처럼-갈라지다, 가뭄-같이-메마르다 등
   b. {홍수}를 포함하는 {NV}유형의 노이즈 예: 홍수-처럼-밀려오다, 홍수-같이-흘러넘치다 등
   c. {지진}을 포함하는 {NV}유형의 노이즈 예: 지진-같이-무너져내리다, 지진-처럼-강렬하다 등

재난재해 어휘+{비교급조사}에 결합하는 노이즈 용언은 대체적으로 재난재해 명사구에 후치되는 특성을 지닌다. 이는 "-처럼/같이" 등과 결합하는 대상 어휘들의 특징을 덧붙이기 위한 것으로 보이며 "가뭄처럼 갈라져서"/"홍수같이 흘러넘쳐"/"지진처럼 강렬한" 등의 표현들이 이러한 표현 유형들에 해당된다. 또한 재해어휘의 명사구에 전치되는 용언의 경우 그 수가 상대적으로 적은 편인데, "(기부가) 메말라 가뭄처럼.."이나 "울면 홍수처럼.."과 같은 형태들을 들 수 있다. 이렇게 추출된 용언들은 각 도메인별 노이즈 패턴 구성 LGG에 삽입된다. {NV} 유형의 노이즈 패턴이 기술된 LGG를 코퍼스에 적용하면 다음과 같은 결과를 획득하게 된다.

{S} 힘을 끌어모으면 지진[DISASTER_NOISE]과 황사도 일으킨다고 합니다.{S}
{S} 지진처럼 붕괴되[DISASTER_NOISE]었다.{S}
{S} 널 지진처럼 휩쓸[DISASTER_NOISE]래.{S}
{S} 숙소 20층인데 지진처럼 흔들린[DISASTER_NOISE]다.{S}

그림 10. {Disaster_Process_지진.grf} 그래프를 통해 추출된 지진 도메인의 노이즈 텍스트

이상에서 특정 자연재해 도메인의 노이즈 태깅을 위해 구성된 그래프들은 하나의 그래프로 통합될 수 있다. <그림 11>은 이 노이즈 패턴 전체를 표상하는 LGG로서, 이 그래프를 코퍼스에 적용하여 추출되는 모든 표현들은 노이즈로 인식되어 태깅됨으로써 별도로 분리 처리되는 것이 가능해진다.

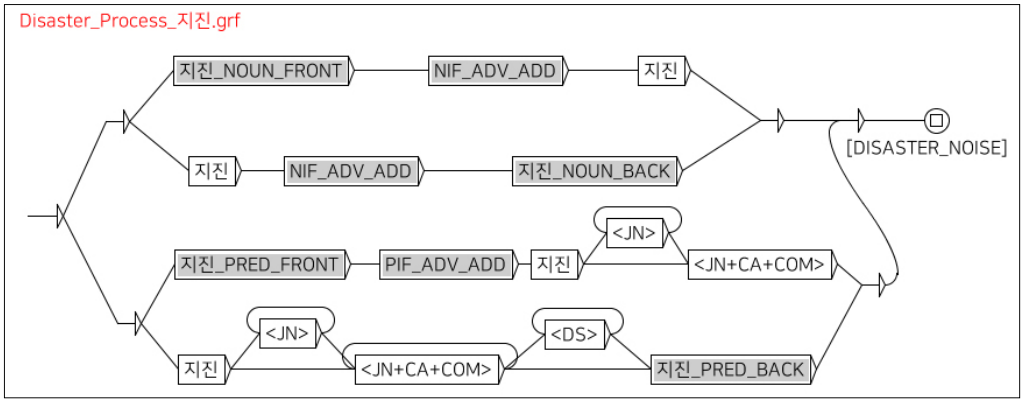

그림 11. 전체 노이즈를 인식하기 위한 {Disaster_Process_지진.grf} 그래프의 구성

4.1.3 자연재해 어휘를 내포한 관용구 노이즈 표현

자연재해 코퍼스의 노이즈에는 속담 및 관용구 유형이 관찰된다. 해당 표현들은 다단어 표현(Multi-Word Expression)의 유형으로써, 어휘 간 결합에 의한 합성성 원리(principle of compositionality)를 준수하지 않고 새로운 의미를 파생시키는 표현 부류를 의미한다. 일반적으로 내포된 자연재해 어휘의 의미가 은유적 표현의 성격으로 변환되므로, 자연재해 현상에 대한 사용자 생성문을 분석하기 위한 대상 텍스트에 해당되지 않는 구문을 구성한다.

본 연구에서는 이러한 유형들을 수집하기 위해 자체적으로 구축한 크롤러를 통해 네이버 사전의 관용구 목록을 수집하였으며, 이들 중 자연재해 대상 어휘 {가뭄/홍수/지진}이 나타나는 표현들을 색인하여 다음과 같은 노이즈 리스트를 추가하였다.

표 4. {가뭄/홍수/지진} 코퍼스의 굳어진(frozen) 표현 노이즈 유형

Index	combination		
	{가뭄} corpus	{홍수} corpus	{지진} corpus
1	가뭄에 콩 나듯	홍수를 이루다	
2	오랜 가뭄 끝에 단비	장마 때 홍수 밀려오듯	
3	뿌리 깊은 나무 가뭄 안 탄다		

4.2 사회재난 {붕괴/충돌/폭발}에 대한 노이즈 처리 리스트 구축

4.2.1 {NN} 유형의 노이즈 패턴

사회재난 어휘와 명사의 결합 유형을 추출하기 위해, 앞서 4.1장의 <그림 2>에서 소개된 것과 마찬가지로, DECO 사전이 적용된 사회재난 대상 코퍼스에서 나타나는 명사 연쇄를 추출하기 위한 LGG 그래프를 구축하였다. 이를 활용하여 각 코퍼스의 대상어휘인 {붕괴/충돌/폭발}의 좌우에 출현 가능한 명사의 리스트를 일괄 추출하였다. 이때 추출을 위해 기술된 명사 부류는 앞서의 경우와 마찬가지로 DECO 사전에 기술된 의존명사와 대명사, 단일음절 명사를 제외한 그 나머지 모든 유형의 명사들로 설정되었다. 이 과정을 통해 추출된 사회재난 어휘의 명사 결합 유형은 각 코퍼스에 따라 다음 표와 같이 분포하였다.

표 5. {붕괴/충돌/폭발} 코퍼스의 "명사 + 명사" 결합 패턴을 구성하는 고빈도 명사기반 리스트

Matching	Occurrences	Matching	Occurrences	Matching	Occurrences
설정 붕괴	944	소행성 충돌	450	멋짐 폭발	1377
정신 붕괴	871	무력 충돌	413	미모 폭발	652
캐릭터 붕괴	722	비탄성 충돌	371	귀여움 폭발	473
멘탈 붕괴	680	설정 충돌	317	광대 폭발	396
붕괴 사고	519	충돌 사고	290	간지 폭발	335
북한 붕괴	468	운석 충돌	260	감성 폭발	331
붕괴 현상	455	군사 충돌	256	매력 폭발	305
밸런스 붕괴	413	정면 충돌	234	분노 폭발	256
체제 붕괴	341	군사적 충돌	208	귀욤 폭발	256
정권 붕괴	256	물리적 충돌	199	인기 폭발	229
방사성 붕괴	244	탄성 충돌	183	흥 폭발	204
붕괴 위험	234	의견 충돌	166	눈물샘 폭발	201
세계관 붕괴	229	대상 충돌	137	감정 폭발	174
건물 붕괴	216	경찰 충돌	112	덕심 폭발	167
왕조 붕괴	204	잠수함 충돌	108	눈물 폭발	153
평양 붕괴	192	낚싯배 충돌	101	귀염 폭발	147
자아 붕괴	156	충돌 가능성	97	케미 폭발	141
붕괴 위기	155	연쇄 충돌	96	카리스마 폭발	116
소련 붕괴	154	충돌 급유선	92	심장 폭발	99
버블 붕괴	131	충돌 우려	80	반응 폭발	93

위의 표를 구체적으로 살펴보면, 사회재난 어휘 {붕괴}와 관련된 명사 결합 유형에는 문학이나 드라마와 같은 컨텐츠에서 자주 사용되는 표현인 "설정 붕괴/캐릭터 붕괴"등이 높은 빈도로 관찰되었으며, 정치 텍스트와 관련된 "체제 붕괴/정권 붕괴"등의 표현들이 뒤를 따랐다. 또한 일상적인 표현으로 자주 사용되는 "정신 붕괴/멘탈 붕괴"와 같은 유형들이 자주 나타남에 따라 이들은 노이즈 리스트로 분류시킬 필요가 있음을 확인하였다.

사회재난 대상 어휘 {충돌}과 같은 경우, 정치권에서 자주 활용되는 "군사 충돌/군사적 충돌/경찰 충돌" 등과 같은 표현들이 고빈도로 나타나는 것을 확인할 수 있다. {폭발}과 관련된 사회재난 코퍼스에서는, "감성 폭발/인기 폭발"과 같이 대상 어휘와 결합하는 상위 빈도의 대부분의 유형이 노이즈로 나타났다. 이러한 사실은 사회재난으로써의 "폭발"과

관련된 텍스트가 사건의 시의성에 따라 일시적으로만 나타나기 때문에, 관련 어휘의 대용량 사용자 생성문 코퍼스를 수집했을 시 해당 유형의 많은 표현들을 노이즈로써 필터링해야 할 필요성을 보여주었다.

앞서 자연재해의 경우와 마찬가지로, 이러한 유형의 고빈도순 리스트에 대한 교차 검증을 위해, DectoTex의 {Expanding Lexica via Word2Vec} 모듈을 사용하여 대상 어휘의 유사도를 관찰하여 이를 노이즈 리스트 구축에 활용하였다. Word2Vec 결과 및 앞서 추출한 고빈도 리스트를 토대로, 각 도메인별 코퍼스에서 나타나는 사회재난 어휘의 명사 결합형에 대한 전·후치 노이즈 명사들을 선별하였다. 이를 <그림 5>에서 보인 그래프에 적용하여 노이즈 리스트를 코퍼스에서 일괄 추출하는 것이 가능하였다. 이렇게 추출된 노이즈 리스트의 일부를 보이면 다음과 같다.

(7) {붕괴/충돌/폭발} 코퍼스의 사회재난 어휘와 명사 결합 유형 노이즈 리스트 일부
   a. {붕괴}를 포함하는 {NN}유형의 노이즈 예: 멘탈-붕괴, 밸런스-붕괴, 붕괴-게임 등
   b. {충돌}을 포함하는 {NN}유형의 노이즈 예: 의견-충돌, 개혁-충돌 등
   c. {폭발}을 포함하는 {NN}유형의 노이즈 예: 분노-폭발, 인기-폭발, 매력-폭발, 폭발-비주얼 등

이와 같은 과정을 통해 획득된 리스트는 앞서 4.1.1에서와 동일한 방식으로 세부 도메인에 따라 각각 LGG 형태로 구성되었으며, 이를 통해 궁극적으로 해당 도메인의 노이즈 패턴을 기술하는 데에 적용되도록 하였다.

4.2.2 {NV} 유형의 노이즈 패턴

비교격 조사와 결합한 사회재난 어휘에 용언이 결합하는 {NV} 유형의 노이즈 후보 추출을 위해, 앞서 4.1.2의 <그림 8>과 동일한 LGG 그래프에 대상 어휘를 사회재난 어휘로 치환하는 방법을 활용하였다. 이러한 그래프를 통해 사회재난 어휘에 "처럼/같은/같이/와 같은/와 같이"가 결합된 형태의 전후에 출현하는 용언 부류를 함께 추출하는 것이 가능하다. 여기서도 앞서와 동일하게 이와 같이 추출된 동사들에 대한 검증을 통해 "사회재난 어휘+{비교격조사}"와 용언의 공기패턴이 구성하는 노이즈 리스트의 구축이 진행된다. LGG에서 기술되는 용언은 실제 코퍼스에서 나타나는 다양한 변이형의 처리를 위해 DECO 사전에서 제공하는 활용형 정보를 활용하였다. 이를 토대로 인식된 활용형은 최종 노이즈 리스트에서 기본형(lemma) 형태로 치환된다. 이러한 방식으로 앞서 자연재해 어휘와 용언이 공기하는 유형과 마찬가지로 사회재난 어휘와 용언이 공기하는 {NV} 유형의 노이즈 리스트가 구축된다. 일부 예를 보이면 다음과 같다.

(8) {붕괴/충돌/폭발} 코퍼스의 '사회재난 어휘+비교격 조사'와 용언 결합 노이즈 리스트 일부
   a. {붕괴}를 포함하는 {NV}유형의 노이즈 예: 붕괴-처럼-무너지다, 붕괴-같이-노출되다 등
   b. {충돌}을 포함하는 {NV}유형의 노이즈 예: 충돌-처럼-굉장하다, 갈등이다-충돌-같이 등
   c. {폭발}을 포함하는 {NV}유형의 노이즈 예: 폭발-처럼-거대하다, 폭발-같이-난리나다 등

이렇게 구성된 노이즈 리스트는 앞서 4.1.2장에서 서술한 것과 동일한 방식으로 LGG 로 구조화된다. 이와 같이 사회재난 관련 키워드에 비교격 조사가 결합할 때에 공기하는 용언 정보들을 활용하여 해당 도메인의 노이즈 표현을 필터링

하여 태깅할 수 있게 하는 언어자원으로 구성된다. 현재 네이버가 제공하는 관용구 사전 목록을 검토한 결과 사회재난 어휘와 관련된 관용 표현은 관찰되지 않아, 이와 관련된 관용구 목록은 추가되지 않았다. 추후 사회재난 대상 어휘의 확장과 함께 이러한 관용적 노이즈 표현들이 보완 검토될 수 있을 것으로 기대된다.

5. 노이즈 필터링을 위한 LGG 문법의 적용 및 평가

5.1 구축된 노이즈 리스트 규모

이상과 같은 과정을 통해 본 연구에서는 자연재해/사회재난과 관련된 일련의 키워드에 대한 노이즈 필터링을 위한 LGG를 구축하였다. 이를 통해 획득된 각 도메인의 재난재해 어휘의 전후에 위치하는 노이즈 리스트 규모는 다음 <표 6>에서 보이는 바와 같다.

표 6. 재난재해 어휘와 관련된 노이즈 리스트의 규모

	자연 재해						사회 재난					
도메인	{가뭄}		{홍수}		{지진}		{붕괴}		{충돌}		{폭발}	
결합 유형	{NN}	{NV}	{NN}	{NV}	{NN}	{NV}	{NN}	{NV}	{NN}	{NV}	{NN}	{NV}
어휘 개수	193	11	266	21	37	11	377	11	486	10	318	14
총계	204		287		48		388		496		332	

노이즈 리스트 구축을 위해 우선 고려한 일정 공기 패턴의 최상위 빈도에서부터 최저빈도수 2까지 동일한 기준을 적용하여 검토하는 작업이 선행되었다. 이는 전체 코퍼스에서 나타나는 재해재난 결합유형 '타입(type)'의 25~30%에 해당하는 것으로, 전체 타입 수를 고려했을 때에 많은 분량을 차지하지는 않는다. 하지만 이를 통해 구축된 리스트의 규모는 각 카테고리 별로 차이를 보였는데, 먼저 자연재해 도메인에서 나타난 노이즈 표현의 총계는 539개, 사회재난 도메인에서 구축된 1,216개로 2배 이상의 차이를 보였다. 이는 자연재해 어휘와 결합하는 노이즈 어휘들이 사회재난에 비해 조금 더 제한적으로 사용될 수 있음은 물론, 특정 유형의 노이즈 표현들만 반복되어 사용되었기 때문이라 판단된다. 특히 자연재해 도메인 {지진}과 같은 경우 이러한 특징이 극단적으로 나타난 것으로 보인다. 그에 비해 사회재난 도메인에서는 보다 다양한 표현들이 노이즈로 나타나고 있다는 점을 알 수 있는데, 사회재난 도메인에서 사용되는 용어들이 언중들의 일상 용어에 다양하게 활용되고 있음을 말해준다. 따라서 이러한 노이즈를 제거하지 못할 경우 실제 사회재난과 관련된 텍스트를 추출하는 데에 많은 어려움이 발생하게 될 것이다.

5.2 DECO 노이즈 필터 프로그램의 사용

구축된 리스트의 적용을 위해서는 한국외국어대학교 DICORA 연구센터에서 자체적으로 개발한 DECO Noise Filter 프로그램을 사용하였다. 해당 프로그램의 외현 및 작동 양상을 간단하게 정리한 순서도는 다음 <그림 12>와 같다.

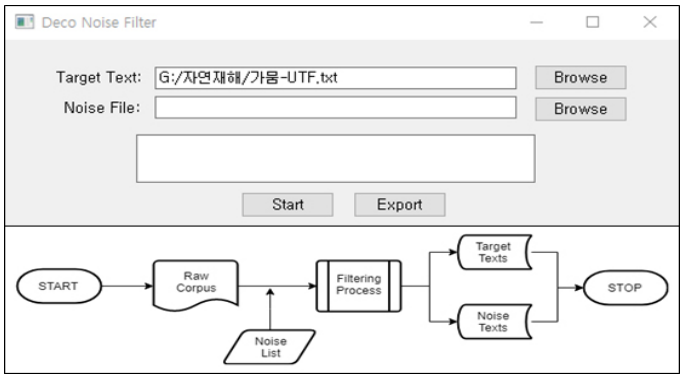

그림 12. Deco Noise Filter 프로그램의 작동 순서도 및 프로그램 구성

<그림 12>의 프로그램은 사용자가 호출하는 노이즈 리스트 파일을 이용하여 해당 표현이 문장 내에서 나타나는 경우 이를 노이즈 텍스트로 인식하여 필터링 처리하는 동작을 수행한다. 프로그램의 실질적인 작동을 위해, 프로그램의 최상단에 위치하는 탭에는 앞서 LGG를 통해 미리 주석된 대상 코퍼스를 호출해야 한다. 두 번째 입력탭인 Noise File 탭에는 일반적인 .txt 파일을 호출해야 하는데, 반드시 그 내부에 "[DISASTER_NOISE]"라는 텍스트를 입력해주어야만 주석 처리를 통한 노이즈 일괄 제거가 가능해진다. 또한 부적절한 광고나 스팸 및 특정 문자열을 제거하고 싶은 경우, 텍스트 파일에 표현들을 입력하는 방법을 통해 제거요소를 확장시킬 수 있다. 정규표현식(regular expression)을 사용하는 경우, 입력하고자 하는 정규표현식 앞뒤에 "**"를 부착하여 이 기능을 활용할 수 있다.

이러한 설정을 통해 각 파일이 입력된 후 "Start" 버튼을 누르게 되면 프로그램은 먼저 입력된 텍스트에서 노이즈 표현이 내포되지 않은 텍스트 파일과 노이즈 표현이 내포된 비대상 텍스트 파일을 서로 분리하여 이들을 각각 별개로 출력해 준다. 이를 통해 아래와 같이 노이즈를 제거하는 처리가 가능해진다.

표 7. {붕괴/충돌/폭발} 코퍼스의 "사회재난 어휘 + 비교격 조사" 유형 노이즈 리스트의 일부

Text Input	
지진 경보 때문에 깜짝 놀라서 핸드폰 떨어트렸어요. 너무 당황스러워서 동공에 지진[DISASTER_NOISE]이 나더라. 가치관의 충돌[DISASTER_NOISE] 때문에 결국 헤어졌다고 하더라구요. 911때 비행기 충돌하던게 아직도 생각나요.	
Text Output	
Target Text	Filtered Text
지진 경보 때문에 깜짝 놀라서 핸드폰 떨어트렸어요. 911때 비행기 충돌하던게 아직도 생각나요.	너무 당황스러워서 동공에 지진[DISASTER_NOISE]이 나더라. 가치관의 충돌[DISASTER_NOISE] 때문에 결국 헤어졌다고 하더라구요.

5.3 실험 및 평가

이상의 과정을 통해 구축된 노이즈 태깅 자원을 시험 및 평가하기 위해, 본 연구에서는 2018년 1월 1일부터 6월 31일까지 각각의 자연재해/사회재난 도메인의 대상 어휘가 나타난 트윗을 추출하였다. 해당 과정을 통해 추출된 트윗 텍스트 중 300개의 트윗을 무작위로 각 도메인에 따라 선별하였으며, 해당 트윗의 재난재해 관련성에 따라 정답지를 작성하였다. 이렇게 작성된 정답지를 Deco Noise Filter 프로그램에서 노이즈 리스트로 적용하여 출력되는 결과를 기반으로 그 성능을 평가하였다. 그 수행 결과는 다음과 같이 나타났다.

표 8. 각 도메인의 500문항 정답지에 대한 성능 결과표

	자연 재해			사회 재난		
코퍼스 도메인	{가뭄}	{홍수}	{지진}	{붕괴}	{충돌}	{폭발}
Precision(%)	98.3	98.2	97.4	97.6	99.2	99.7
Recall	88.7	90.1	89.9	66.5	79.2	82.9
f-measure	93.2	93.9	85.8	79.1	88.0	90.5

<표 8>의 평가 결과에서 모든 도메인에서의 노이즈 텍스트 분류와 관련된 정확률(precision)이 98.4의 평균값을 보이는 것을 확인할 수 있다. 노이즈 리스트가 비교적 정확하게 작성되었음을 알 수 있다. 재현률(recall)의 경우 자연재해와 관련된 노이즈 분류에서는 평균 88.2로 나타났으나, 사회재난 도메인의 경우 76.2로 비교적 낮은 재현율이 나타나는 것으로 확인되었다. 이러한 차이가 발생하는 이유는 자연재해 어휘에 비해 사회재난 어휘의 경우 더 다양한 체언 및 용언 부류와 결합을 통해 노이즈 표현을 생산하고 있기 때문으로 추정된다. 즉 은유적 표현을 생성해내는 관점에서 자연재해 어휘보다 사회재난 어휘가 더 개방적인 특성을 가지고 있다고 할 수 있다. 따라서 사회재난 코퍼스와 관련된 노이즈 리스트를 구축하는 경우, 상위 빈도뿐만 아니라 비교적 낮은 빈도로 등장하는 표현들에 더 많은 관심을 두어야 할 것으로 예측된다. 또한 추후에는 언어학적 패턴을 통해 추출한 다양한 노이즈 용례들이 코퍼스에서 단 한번만 나오는 표현(hapax legomena)로 등장하는 경우도 고려하여 극저빈도로 나타나는 다양한 타입(type)의 용례들을 최대한 반영하는 것이 필요할 것으로 생각된다. “지진정보”의 잘못된 표현인 “지진정”과 같이 일부 미분석 어절로 나타나는 표현들의 경우는 실제 재난재해와 관련이 없기 때문에 이러한 표현들 또한 리스트 확충 시에 재고되어야 할 유형으로 보인다.

6. 결 론

본 연구에서는 재난재해와 관련된 코퍼스를 수집하고 대상 텍스트를 분류해내는 작업에 있어, 여기에 포함되는 노이즈 유형의 텍스트를 필터링하기 위한 언어학적 방법론을 소개하였다. 해당 방법론을 통해 재난재해 코퍼스 도메인에서 나타나는 각 노이즈 표현들을 태깅 및 주석하기 위한 언어 자원을 LGG 그래프문법으로 구축하였으며, 해당 자원의 유용성을 평가하기 위하여 평가셋을 구축, 그 적용 결과의 성능을 확인하는 과정을 통해 그 유용성을 입증하였다.

재난재해 코퍼스에서 나타나는 노이즈 제거 연구는 해당 도메인의 코퍼스에서 나타나는 노이즈 표현들이 상당히 방대함에도 불구하고 국내외 관련 연구가 활발히 진행되지 않았다. 이는 그 필요성에 비해 중요도에 대한 인식이 상대적으로 부족했기 때문으로 보인다. 코퍼스 기반 연구에서 텍스트의 노이즈를 최소화하는 작업은 데이터의 질과 직접적으로 관계되기 때문에, 정확한 연구 결과를 산출해내기 위해서는 반드시 노이즈 제거 과정을 거쳐야 한다. 본 연구에서는 해당 노이즈 리스트를 추출하기 위해 언어학적인 용례 추출을 중심으로 Word2Vec 모듈의 출력 결과를 참고하는 노이즈 추출 방법론을 제시하였으며, 해당 방법론을 통해 구축한 노이즈 표현의 재난재해 도메인별 LGG와, 해당 노이즈 태그를 접목시켜 필터링하는 응용프로그램을 개발 및 제시하였다.

과학 기술의 비약적인 발달이 이루어진 현대 사회에서도, 인간은 지진이나 해일과 같은 자연재해를 비롯하여 시시각각 벌어지는 충돌, 오염, 폭발과 같은 사회재난의 위협으로부터 안전하지 못하다. 이에 따라 어떤 재난적 사건에 대한 언중의 감정이나 의견을 읽어내어 대응 및 대비책을 마련하기 위해 재난재해와 관련된 사용자 생성문의 텍스트 분석이 상

당히 유용하게 활용될 수 있다. 하지만 데이터 분석 이전에 대상 텍스트에 노이즈 표현이 대량으로 존재한다면, 해당 데이터 분석의 신뢰도는 심각하게 저하될 것이다. 본 연구에서는 재난재해 텍스트에서 보다 정확한 대상 텍스트를 분류해 낼 수 있도록 노이즈 필터링을 위한 언어자원 구축 방법론을 제안하였다. 이를 토대로 재난재해 텍스트에서 나타나는 노이즈 표현에 대한 연구가 추후 다양한 영역으로 확장되어, 향후 보다 세분화된 분야에서 이러한 전처리를 통한 양질의 언어 데이터가 획득될 수 있기를 기대한다.

참고문헌

고아라. 2013. 조사 "같이"와 "처럼"의 의미와 기능에 대한 연구.『건지인문학』9, 5-30.

김진해. 2014. 은유적 합성명사의 결합관계와 인지언어학적 해석.『국어학』70, 29-57.

남지순. 2018.『코퍼스 분석을 위한 한국어 전자사전 구축방법론』. 도서출판 역락.

박태연· 한희정· 김용· 김수정. 2017. 재난안전정보의 통합 관리를 위한 분류체계 현황분석 및 개선방안에 관한 연구.『한국비블리아학회지』28.3, 125-150.

신봉희· 전혜경. 2018. 빅 데이터를 이용한 재해 정보 지원에 관한 연구.『한국융합학회논문지』9.8, 25-32.

신자행. 2016. 빅 데이터 기반의 재난정보관리 방안. 서울대학교 대학원 석사학위 논문.

안갈승· 서민지· 허선. 2017. 효과적인 산업재해 분석을 위한 텍스트마이닝 기반의 사고 분류 모형과 온톨로지 개발.『한국안전학회지』32.5, 179-185.

임채훈. 2002. 국어 비유구문의 의미연구-"처럼", "만큼"을 중심으로-.『한국어 의미학』10.

유광훈· 남지순. 2017. DecoTex Users' Manual. DICORA-TR-2017-12. DICORA, 한국외국어대학교.

황창회· 남지순. 2018. SNS 사용자 생성문에 대한 코퍼스 수집 시스템 소개: Deco Crawlers. DICORA-TR-01-2018. DICORA, 한국외국어대학교.

Baek. S., H. Jeong, and K. Kobayashi. 2013. Disaster Anxiety Measurement and Corpus-Based Content Analysis of Crisis Communication. *IEEE International Conference on Systems, Man, and Cybernetics* 1789-1794.

Gross, M. 1997. The Construction of local grammars. *Finite-State language processing*, Roche & Schabes (eds.), the MIT Press.

Gross, M. 1999. Nouvelles applications des graphes d'automates finis à la description linguistique, *Lingvisticae Investigationes* 22.1-2, 249-262.

Matherson, D. 2018. The performance of publicness in social media: tracing patterns in tweets after a disaster. *Media, Culture & Society* 40.4, 584-599.

Paumier, S. 2003. *De la reconnaissance de formes linguistiques a l'analyse syntaxique*, Ph.D. dissertation, Univ of PEMLV in France.

황창회(제1 저자), 대학원생
경기도 용인시 처인구 모현 외대로 81
한국외국어대학교 언어인지과학과
E-mail: 201830252@hufs.ac.kr

남지순(교신 저자), 교수
경기 용인시 처인구 모현 외대로 81
한국외국어대학교 언어인지과학과
E-mail: namjs@hufs.ac.kr